対話データの再帰結合神経回路による学習と相槌タイミング予測 ～音声特徴と視線特徴の影響～

○佐野正太郎 西出俊 奥乃博 尾形哲也 (京都大学大学院)

1 はじめに

近年, ロボット対話システムや音声対話エージェントなど, 人間と機械がコミュニケーションを行うことのできるユーザインタフェースの実現をめざして各所で研究が行われている.

しかし, 現在の一般的なシステムでは, ユーザとシステムが一律な決まりのもとで交互に発話することだけが仮定されている. これは, 対話システムにユーザが馴染めない大きな要因となっている. 実際の人間どうしの対話では, 自然なタイミングでの相槌や話者交代が行われており, 聞き手側の話の理解状況が話し手にフィードバックされている. 対話システムにおいても, 相槌のようなフィードバックが得られれば, ユーザにとってより使いやすいシステムになると考えられる. とくに相槌はユーザからの入力をスムーズに受け取る上で重要な役割を占める.

そこで, 本稿ではロボットやエージェントが適切なタイミングで相槌を打てるシステムの開発を目指した. 対話コーパスから自然な相槌タイミングを学習するための新たな手法を提案し, ポスターセッション対話を対象とした性能評価実験を行った.

2 従来研究

相槌のタイミングに関する研究は, これまでにも様々に行われている. 対話における聞き手の相槌タイミングは, 話し手の言語情報と韻律情報に影響されることが知られている [1]. 特に韻律を用いた研究では, 発話が終了する段階に相槌が集中することに着目しており, 発話ポーズを合図として, 発話終了を検出した上で相槌のタイミングを判定している.

このような手法には 2 つの問題点がある. 1 つに, 相槌は必ずしも話し手の発話ポーズに付随して現れるわけではない. 話し手の発話にオーバーラップする相槌に焦点を当てた研究は少ない. もう 1 つの問題として, 従来のシステムは話し手の発話ポーズに対して事後的に相槌を挿入しているが, 相槌は発話ポーズよりやや早い段階から被せるように発生し始める傾向にある.[2] したがって, 相槌の有無は発話ポーズ以前から予測的に判定されることが望ましい.

一方で, 相槌をはじめとする会話現象は動的な生成行為であることから, 対話中の引き込み現象などに基いた力学モデルの有効性が指摘されている [3]. 本研究では, この点に着目し, 神経力学モデル Multiple Timescale Recurrent Neural Network (MTRNN)[4] による相槌タイミング決定システムを考案した. このシステムでは, 相槌のタイミングを予測的に決定することができる. また, 発生時点をポーズ点に限定しないタイミング決定が可能となる.

3 タイミング予測モデル

本システムでは音声情報と映像情報を入力とし, 相槌タイミングの予測値を出力する (予測値については 4.1.4 節で述べる). 入力された音声と映像は特徴量抽出器を通して F0, 音声パワー, 視線情報などに変換される. これらのデータを予測学習器 MTRNN に入力し, 相槌の発生タイミングを得る.

MTRNN は現在の状態を入力に, 次ステップの状態を出力する予測器である. 階層構造を持ったネットワークであり, 入出力を司る IO ノード, IO ノードより遅れて変化する Cf ノード, 更に遅れて変化する Cs ノードを持つ. 内部状態を階層的に変化させることで, 複雑な時系列予測が可能となる. 本研究で実験に用いた MTRNN は, 特徴量に対応した 2～5 個の IO ノード ($IO_{parameter}$), 相槌タイミングに対応する 1 個の IO ノード ($IO_{backchannel}$), 15 個の Cf ノード, 8 個の Cs ノードから構成される. $IO_{parameter}$ を除く各ノードは同種のノードへのフィードバックループを持つほか, Cf は IO, Cs と相互に結合されている. また, $IO_{parameter}$ には外部から特徴量が入力される. 時刻 t におけるノード i の値 $u_{t,i}$ は以下のように求める.

$$y_{t,i} = \frac{1}{1+exp(-u_{t,i})} \tag{1}$$

$$u_{t,i} = \begin{cases} 0 & \text{t=0} \\ (1-\frac{1}{\tau_i})u_{t-1,i} + \frac{1}{\tau_i}\sum_j \omega_{ij} x_{t,j} & \text{otherwise} \end{cases} \tag{2}$$

$$x_{t,j} = \begin{cases} g_j(t) & \text{j} \in IO_{parameter} \\ y_{t-1,j} & \text{otherwise} \end{cases} \tag{3}$$

ここで, $g_i(t)$ は時刻 t における特徴量 i の値, τ_i はノード i の時定数, ω_{ij} はノード j からノード i への結合重み, $y_{t,i}$ は時刻 t におけるノード i の出力である. また, 時定数 τ_i は IO で 2, Cf で 5, Cs で 70 とした. 時定数が大きいほどノードの状態は緩やかに変化する. 重みの学習は Back Propagation Through Time 法 (BPTT 法)[5] によって行った.

4 評価実験

提案モデルについての評価実験を行った. 実験ではポスターセッションのコーパスを対象とし, プレゼンタの韻律と視覚特徴量から聞き手 1 人の相槌タイミングを再現した. 実験には角らによる IMADE プロジェクト [6] の中で収集, 分析されたデータのうち, ポスターセッションをタスクとした 20 分の対話データを用いた.

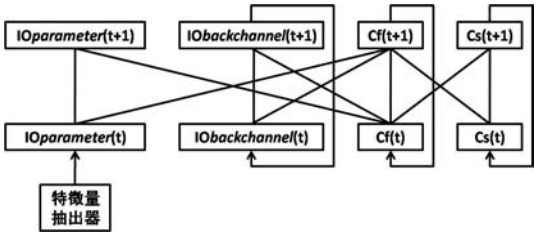

図 1MTRNN による時系列生成モデル

ここでは 2 つの評価実験を行っている. 1 つ目の実験では提案モデルの性能を定量的に評価した. 2 つ目の実験では特徴量を変化させた際の性能の違いを比較した.

以下, 使った特徴量について, 2 つの実験の手順について, それぞれ詳細を述べる.

4.1 利用情報の候補と取得方法

特徴量の候補として, “F0”, “音声パワー”, “視線が聞き手に向くタイミング”, “視線が対象物に向くタイミング”, “頷くタイミング” をプレゼンタの行動から抽出した. 特徴量の条件として重視した事は, “相槌との関連性が強いと考えられること” および “逐次取得可能な情報であること” の 2 点である. したがって, 言語情報など逐次取得が難しい情報は使用していない.

4.1.1 F0, 音声パワー

相槌タイミングの決定にあたっては韻律情報の有効性が確認されている [1]. 我々のシステムにおいても F0 と音声パワーを特徴量とした. F0 とパワーの分析周期は 100ms とし, F0 の検出には自己相関関数を用いた.

4.1.2 視覚情報

我々はいくつかの視覚情報も利用した. そもそも会話はマルチモーダルな行為であり, 音声のほかにも, 視線, ジェスチャ, 表情といった様々な要因が作用している. したがって, 相槌のタイミングもこれらの影響を受けている可能性が考えられる.

用いた視覚情報は “視線が聞き手に向くタイミング”,“視線が対象物に向くタイミング”,“頷くタイミング” の 3 つである. これらの情報は, IMADE プロジェクトにおいて, 予め分析されていたものを用いた. 視線に関する情報は, 参加者の頭部に備えられた方向センサによって抽出されていた. 頷きの抽出は, 同じく参加者の頭部に備えられたセンサが使われており, 頭部の上下運動から判定されていた. これらの情報はアノテーションツール iCorpusStudio[8] によってラベル付けされていた.

4.1.3 相槌タイミングの取得

相槌のタイミングは人手によって分析されていた. この情報も, 視覚特徴量特徴量と同様に iCorpusStudio によってラベル付けされていた.

4.1.4 MTRNN 用学習データへの変換

MTRNN は時系列データを生成するモデルである. 相槌のような事象も時系列データとして扱わなければならない. また, その値は連続的であることが望ましい. そこで, 以下の式で示される, 相槌の予測値 $f(t)$ を定義し, 相槌を連続値化した.

$$f(t) = \max_{n\in[1,N]} [\exp\{-\frac{(t-t_n)^2}{2}\}] \tag{4}$$

N は観測データにおける相槌の総数, t_n は n 回目の相槌が発生した時刻である. 視覚情報のデータもアノテーションされたラベルの段階では離散的なデータであるため, 次に示す式で特長量 i に対する連続的な時系列 $g_i(t)$ を定義した.

$$g_i(t) = \max_{n\in[1,N_i]} [\exp\{-\frac{(t-t_{i,n})^2}{2}\}] \tag{5}$$

4.2 実験手順

コーパス中の聞き手の相槌を含む部分区間 (5 秒から 15 秒ほど) を計 44 個用意した. うち 22 個を学習用データとして用い, BPTT 法によってネットワークの結合重みを決定した. 残りの 22 個は評価用データとし, 2 章で述べた手法に従い相槌タイミング時系列を生成した. 学習用データ 22 個中には合計で 43 個の相槌が, 評価用データ 22 個中には合計で 35 個の相槌が, 合計で 78 個の相槌が含まれていた. また, それらの相槌のうち 47 個は話し手の発話終了点を含む区間で打たれた相槌, 残る 31 個はそれ以外の時点で打たれた相槌であった.

また, 2 つの実験では実測データの相槌を正解の相槌と定め, 本手法の性能を定量的に評価した. ここで, 正解の判断には相槌の予測値を用いた. システムの生成した相槌の時系列において, 予測値が 0.5 を超えた場合に, システムが相槌を打ったと判断し, この条件を満たす時点にラベル付けした. 同様に, 実測データにおいても相槌の予測値が 0.5 を超える時点にラベル付けした. 実測データにおいてラベル付けされた各時点に対し, その前後 300ms に注目し, この範囲に生成時系列でのラベル点があった場合に, 相槌を正解とした. ただし, 注目した範囲に生成時系列のラベルが 2 個以上存在した場合, その中の 1 つだけを正解とした. これらのラベル個数から計算される再現率, 適合率, F 値によってシステムの性能を評価した.

4.2.1 実験 1:提案システムの性能評価

1 つ目の実験では, 4.1 節で挙げた 5 つの特徴量全てを用いたときの性能を定量的に評価した. 学習の際, MTRNN では結合重みの初期値によって最終的に得られる結合重みが変化する. このため, 結合重みの初期値の組み合わせをランダムに 10 個用意し, 同様の実験を 10 回行った. 学習で得られた 10 個のモデルそれぞれに対して再現率, 適合率, F 値を評価し, F 値における最良モデルを選出した.

4.2.2 実験 2:特徴量による性能比較

特に視覚特徴量に対して, それぞれの特徴量の必要性を確認するため, 特徴量の組み合わせによる性能の変化を比較した. 具体的には, (1) 韻律情報のみを用いた場合, (2) 韻律情報と “視線が聞き手に向くタイミング” を用いた場合, (3) 韻律情報と “視線がポスターに向くタイミング” を用いた場合, (4) 韻律情報と “頷くタイミング” を用いた場合, (5) 全ての特徴量を用いた場合, の 5 通りについて実験 1 と同様の手順で性能を評価し, それ

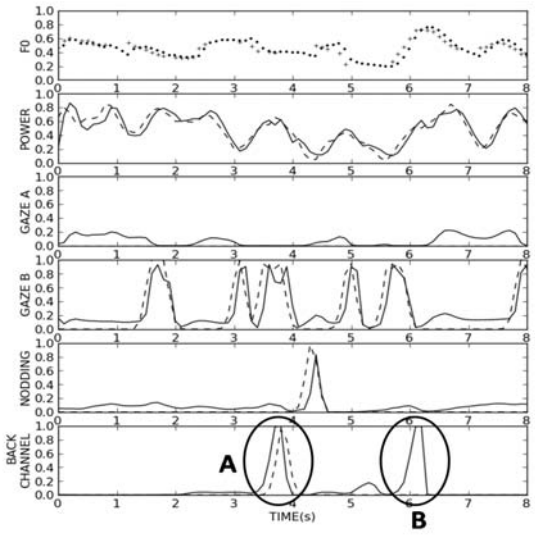

図 2IO ノードの予測結果

ぞれを比較した. それぞれの組に対して, 学習初期値を変化させて 10 回の学習と評価を繰り返し, その平均的な性能を F 値によって評価した.

5 結果

実験 1 では 10 個のモデルのうち, 最良モデルの学習用データにおいては再現率 46.5%, 適合率 52.6%, F 値 49.4%が得られた. 評価用データにおいては再現率 37.1%, 適合率 31.7%, F 値は 34.2%が得られた.

また実験 2 における, F 値の比較を図 4 に示す. 韻律のみを用いた場合に F 値 9.9%, 韻律に加え “視線が聞き手に向くタイミング” を用いた場合に F 値 18.2%, “視線がポスターに向くタイミング” を用いた場合には F 値 14.3%, “頷くタイミング” を用いた場合には F 値 10.9%であった. これらの特徴量を全て用いた場合には F 値 22.2%が得られている.

6 考察

ここでは, 定量的評価と特徴量の比較について考察するとともに, 本手法が相槌の予測を事前に行えているか, ポーズ点以外の相槌に対して十分な性能を示したか, について述べる. また, システムがどのようなネットワーク構造をもって相槌タイミングを判断しているかについても考察する.

6.1 相槌予測への有効性

評価用データに対する生成結果の一例を図 2 に示す. また, その対話ログを図 3 に示す. 図 2 において, 1 段目から 5 段目は特徴量の時系列, 6 段目は相槌タイミングの時系列である. 破線は現実に観測されたデータを表し, 実線は MTRNN が生成した時系列を表す (ただし F0 は実測値を十字点で, 生成値を丸点で表している).

図 2 の 6 段目において破線では 1 つのピークがたっており, この極大値が実際に相槌が打たれたタイミングである. 実線では 2 つのピークが立っており, A の部分では破線のピークと重なっている. つまり, このピークに

1	P	ユーザが興味を示す/ ような箇所を/ とれれば/
2		(0.2)
3	P	有用なんじゃないか/ というような/
4		(0.3)
5	P	え:/ スタンスで/
6	W	ふうん
7	B	え: やっています/
8		(0.5)
9	A	で さっき言ったように/ その/

図 3 対話ログの例

対応する相槌に関しては, 実測のタイミングと MTRNN が挿入したもののタイミングが同時であったといえる. 一方で, B の実線ピークに対しては, 相槌の実測はなく, ここではコーパスになかった相槌が挿入されていた.

ピークは極大値に到達する 0.5 秒以上前から立ち始めていることに注目しておきたい. これは暗に, MTRNN が相槌タイミングの到来を, その少し前から予測していることを示している. 本手法が相槌タイミングの予測に有効であることが, 視覚的に確認できる.

6.2 定量的評価に対する考察

実験 1 では, 未知のデータに対しても, およそ 3 分の 1 程度の割合で元データとタイミングが一致していることがわかる.

ここで, 相槌のタイミングに関しては絶対的な正解の基準は無いため, コーパスに無いタイミングで予測された相槌であっても不自然なものであるとは限らない. 例えば, 図 2 において, B の実線ピークに対応する相槌は, コーパス上に存在しなかったものである. しかし, この領域は対話ログにおいて, 8 行目のポーズが現れる箇所と重複しており, また, 特徴量の F0 が低くなる領域とも重なっている. したがって, この位置での相槌は適切であったとも考えられる.

このようなデータも総合して評価するため, 今後は, 主観的な評価を行う必要がある.

6.3 ポーズに関係しない相槌の評価

正解した相槌のうち, 話し手の発話終了点で打たれた相槌は平均すると 15.9 個で, 発話終了点における全ての相槌 47 個のうち 33.8%の相槌を再現していた. 一方で, その他の点において正解した相槌は平均すると 8.4 個であり, もとの 31 個の相槌のうち 27.1%の相槌を再現していた. いずれの点においてもほぼ同様の性能が示されており, 本手法が発話終了点以外の相槌にも有効であることが示された.

6.4 特徴量の比較

図 4 より, 学習用データ, 評価用データのいずれにおいても, 特徴量として韻律のみを用いる場合よりも視覚情報 1 つを加えた場合の方が F 値が高く出ている. 全ての視覚情報を加えることで, F 値は更に増加している. このことから, 今回用いたいずれの視覚情報も相槌タイミングの予測において有効であることが確かめられた.

6.5 Cf 空間の解析

学習を経た MTRNN が内部にどのような構造を獲得しているかを調べた. 特に生成時の Cf ノードの推移を

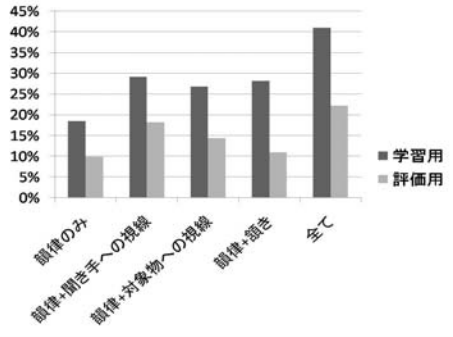

図 4 特徴量による F 値の比較

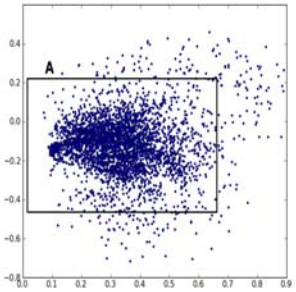

図 5Cf ノード値の分布

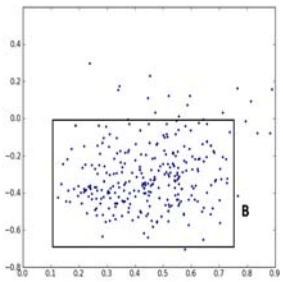

図 6 相槌発生時の分布

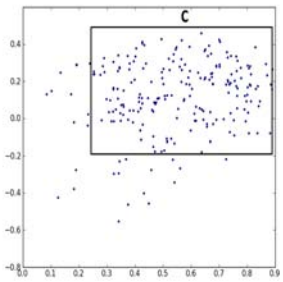

図 7 視線変化時の分布

解析したところ, 視線変化や相槌発生などの事象が, Cf 空間上で特徴的にマッピングされていた.

韻律情報と “視線が聞き手に向くタイミング” を特徴量としたときに平均的な性能を示したモデル 1 つを対象として, 生成時の Cf ノード値の分布をプロットした (図 5). この図では実験に用いた 44 個のデータに対し, 全てのステップにおける Cf ノード値がプロットされている. なお, Cf 空間は本来 15 次元であるため, ここでは 15 次元から 2 次元への主成分分析を行っている. それぞれの寄与率は 36.2%(横軸) と 21.5%(縦軸) である.

ここでは, 図中の A で囲っている領域に値が集中している. 生成におけるほとんどのステップでは, Cf ノード値がこの領域に留まっていることがわかる.

一方で, 図 6 は相槌生成が起こった直後 5 ステップに限定して Cf ノード値をプロットしたものである. ここでは図中 B で囲った部分に値が集中している. 先ほどの A とは分布が大きくずれていることから, B の領域では相槌生成に特徴的なマッピングが行われているとわかる.

図 7 は視線変化の直後 5 ステップに限定した Cf ノード値をプロットしたものである. ここでも図中 C で囲った部分に特徴的な分布が見られる. Cf ノード値は視線変化に対応してこの部分に推移していることがわかる.

以上のことから, 学習によって獲得されたネットワーク構造は,“視線変化” や “相槌” などの事象を強く反映したものであるといえる. これは, 得られたモデルが相槌に関するある種の会話構造を獲得していることを示唆している.

今後, このような構造を更に解析することで, マッピングされた事象と相槌生成の関係について新たな知見を得ることも期待できる.

7 おわりに

本稿では, 神経力学モデル MTRNN による相槌の挿入手法を提案した. 評価実験により, 相槌の発生をそのしばらく前から予測できること, 相槌の発生箇所を発話ポーズなどに限定せず予測できることが確認された. また考察では, 相槌の挿入にあたって視覚情報が重要であることが確認し, ネットワークがある種の会話構造を反映することを述べた.

今後は提案手法をロボット対話システムに実装し, 実環境において提案手法が有効であるかを検討すると同時に, 性能の主観的な評価を行う予定である.

謝辞 本研究の一部は JST さきがけ, 科研費基盤 (S)(B), 科研費学術創成, GCOE の支援を受けた。

参考文献

[1] N. Ward: “Prosodic features which cue back-channel responses in English and Japanese”, Jounal of Pragmatics 32, pp.1177-1207(2000).
[2] Y. Okato, K. Kato, M. Yamamoto and S Itahashi: “Insertion of Interjectory Response Based on Prosodic Information”, Interactive Voice Technology for Telecommunications Applications, 96(1996).
[3] H. Ogawa, T. Watanabe: “InterRobot: a speech driven embodied interaction robot”, Robot and Human Interactive Communication, 2000, pp.322-327(2000).
[4] Y. Yamashita, J. Tani: “Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model: a Humanoid Robot Experiment”, PLoS Comput. Biol, Vol.4(2008)
[5] D. E. Rumelhart, G. E. Hinton, R. J. Williams: “Learning internal representations by error propagation”, Parallel distributed processing: explorations in the microstructure of cognition, vol.1 (1985)
[6] http://www.ii.ist.i.kyoto-u.ac.jp/IMADE/
[7] Y. Sumi, M. Yano, T. Nishida: “Analysis environment of conversational structure with nonverbal multimodal data”, The Twelfth International Conference on Multimodal Interfaces and the Seventh Workshop on Machine Learning for Multimodal Interaction (2010).
[8] http://www.ii.ist.i.kyoto-u.ac.jp/iCorpusStudio/feature.html